Working with Data Final Project

Project Goals

This class will culminate in a group project and presentation. This project will give you experience

- asking interesting research questions
- finding the best data to answer those questions
- contextualizing research using summary statistics and visualizations
- analyzing data with the purpose of learning something specific
- working in a group
- presenting and writing in English

Your project should be one big narrative (or many small ones) – it **should not** just be a collection of code and figures. That is, your Jupyter Notebook must include exposition that ties all of the data, figures, and analyses together into a focused and clear report on your chosen topic(s).

Project Details

For this project, you will work in groups of 3 or 4 people. Groups can choose one of two options for their project.

Option 1: Country Report

Pick a country and use data to create a report on that country. Which aspects of the country you explore is up to you as long as data can be used to do so. Some potential ideas:

- Tell a story about how that country has changed over time (e.g. measures of poverty, economic growth, political participation, etc.)
- Go in-depth on a single topic in that country (e.g. the lumber industry in Canada)
- Focus on specific subregions of that country or compare subregions to each other (e.g. how did economic development vary across Chinese provinces?)
- Illustrate the relationship between variables in the context of that country (e.g. what is the relationship between a person's race and their probability of incarceration in the United States?)
- Perform a predictive analysis in that country (e.g. how well does rainfall predict agricultural yield?)

Option 2: Country Comparison

Pick two (maybe three) countries and compare and contrats them. The comparisons should be interesting, so it will help to choose countries that have some similarities. For example, you could compare countries that

- are geographically close
- have similar industries, economic compositions, or levels of development
- have similar cultures, histories, etc.

You are bound to find some differences between any two countries, but if you find only differences, then the differences you find aren't as interesting! With that in mind, what aspects of the countries you choose to compare is up to you and your group. Topics that work for Option 1 would also work for this project as long as there is a comparison to be made.

Choose Your Own Topic!

While you must focus you analysis around one or more countries, the specific ideas provided on this document are just examples. Your ideas may be completely different or be some combination of the aforementioned ideas. In fact, it is unlikely that a single idea above will produce enough content for the whole project, especially since your project should implement a few of the techniques we cover in class.

Ideally, you can weave many topics into one cohesive narrative, but the topics don't necessarily have to be related – each one can be its own self-contained story. It may be easier to choose many smaller questions or topics than one or two very in-depth ones but that will depend on context. For example, maybe you describe country-wide poverty in one part of the project and analyze a particular city's traffic data in another segment. In this circumstance, poverty and traffic are likely unrelated and that is ok.

Let the Data Guide You

Unfortunately, you cannot answer questions or discuss topics unless you have access to the relevant data. Instead of trying to come up with the perfect question before looking at the data, come up with a few countries that you are interested in learning about, and explore what data is available for those countries. Then, consider what kinds of questions you can answer with that data. In general, it will be harder to find data for smaller, less-developed countries.

A good place to start looking for data is here. Google is also very helpful.

Country Restrictions

If you choose Option 1, you cannot choose The United States, Canada, or China as your country. You also cannot choose the home country of someone in the group. If you choose Option 2, one country you discuss must obey the aforementioned restrictions for Option 1. There are no restrictions on the other countries you use for comparison. Finally, two groups cannot choose the same country for Option 1 or two of the same countries for Option 2. Otherwise, the presentations might overlap significantly. The instructor will make sure this does not happen.

Logistical Details

Project Deliverables & Dates

Each group will submit

- a one page project proposal that details their topics/questions, data sources, and planned approach (due the beginning of class on July 28th)
- an **executed** Jupyter notebook that contains their entire report including code, figures, and exposition (due before class on August 8th)
- a README.md file detailing each group member's name and their contribution to the project
- a GitHub Repository containing the **executed** Jupyter Notebook, data files (that obey GitHub's size limits), a README.md file, and any other relevant files (due before class on August 8th)

Presentation

Groups will also prepare a 12-15 minute presentation on their country which they will give on August 8th. Groups should use their submitted Jupyter Notebook as a visual aid. Do not, however, simply read off of the Jupyter notebook. Even when presenting the exact same information, a presentation should use different words than a report. **Each group member must present for at least 3 minutes.**

These are the only deadlines, but I highly encourage students to communicate with me about the state of their project if they need guidance or an opinion. I will be actively engaging with students during the labs to see how their projects are coming along and answering any questions. Students can also see me after class or email me to schedule appointments.

Grading

The Jupyter Notebook is worth 30% of your final grade, the presentation is worth 15%, and the proposal is worth 5%. Below, I describe the hallmarks of a good project and what I look for below.

Project Grading

A successful project will

- have an introduction, data description, descriptive statistics, visualizations, analysis, and conclusion
- have clear, well-defined goals and accomplish them
- integrate code and figures with text to create a cohesive report
- will reflect effort (many analyses, difficult analyses, custom data sets, well-researched background, well-written, strong aesthetics, etc.)
- will include a data wrangling component
- will include a map
- will include a regression or classification component

Presentation Grading

A successful presentation

- demonstrates effort and preparation
- gives the audience a good understanding of what lies in the report
- uses the Jupyter Notebook effectively as a visual aid but not as a crutch
- involves all group members participating equally
- engages the audience

Proposal Grading

Once I accept a proposal as your project, you get full marks, but I will only accept the proposal if I can understand what your project entails from reading it. If I can't understand or I think you should pursue another topic, you may have to meet with me. Once we have met and I understand your project, you will still get the full 5%.

English

English grammar and spelling will never be graded. We are focused on coding and data analysis in this class. That being said, I do need to understand what you write and what you say, so if your English is not as strong, it might help to have someone in your group who is more proficient in English. I also do not mind if you use chatGPT or another LLM to **edit** your writing for you. Please review what it writes as it can make up things that sound good.

Your Obligation to Your Group

As previously mentioned, it will be difficult to do a good job on this project if you only work in class. Whether you split up the work or try and work on everything together, you will have to communicate and meet with your group outside of class. Please make yourself available to your group members because they will be counting on you to do your share of the work. Do not ghost your group or just work on the project during class while they handle the rest. **If you fail to communicate with your group or contribute to your project, you can and will receive a worse grade on the project.**

FAQ

Why Groups?

Coding and research are frequently collaborative endeavors. Consequently, it is important to learn how to work and communicate with other people in a coding and research context. This will also give you experience collaborating on GitHub with other people instead of just using it as a fancy Drop Box.

Why Countries?

There is a lot of data on the Internet. This class also takes place over a short period of time. On top of all of that, students are simultaneously learning the tools they need to complete the project while working on the project. I do not want students taking a long time to find a topic and data. I also do not want students choosing a Kaggle dataset and doing the same analysis that has been done 100 times before, and I believe this restriction will result in both faster topic selection and more interesting projects. I also think this type of project will afford students more opportunities to apply what they have learned in the class.

Why Country Restrictions?

The universally excluded countries (China, Canada, and US) will be very popular choices with easily available data. Since not everyone can choose these countries, allowing them will result in me picking winner and losers – those who get get to research these countries will have an easier time finding data.

Also, want students to learn about the country they choose by going through this exercise. If they are from that country, they will likely not learn much. This is why I have also excluded doing a report on one of your group member's home country.